Journal of Biomedical Informatics 41 (2008) 904–913

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Automated comparative auditing of NCIT genomic roles using NCBI

Barry Cohen [a,*], Marc Oren [a], Hua Min [b], Yehoshua Perl [a], Michael Halper [c]

[a] Computer Science Department, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA
[b] Fox Chase Cancer Center, Philadelphia, PA 19111, USA
[c] Computer Science Department, Kean University, Union, NJ 07083, USA

ARTICLE INFO

Article history:
Received 7 August 2007
Available online 28 March 2008

Keywords:
NCI Thesaurus
NCBI GenBank
NCBI Entrez Gene
Gene hierarchy
Biological Process hierarchy
Gene terminology
Automated auditing
Web crawler
Biomedical knowledge base

ABSTRACT

Biomedical research has identified many human genes and various knowledge about them. The National Cancer Institute Thesaurus (NCIT) represents such knowledge as concepts and roles (relationships). Due to the rapid advances in this field, it is to be expected that the NCIT's Gene hierarchy will contain role errors. A comparative methodology to audit the Gene hierarchy with the use of the National Center for Biotechnology Information's (NCBI's) Entrez Gene database is presented. The two knowledge sources are accessed via a pair of Web crawlers to ensure up-to-date data. Our algorithms then compare the knowledge gathered from each, identify discrepancies that represent probable errors, and suggest corrective actions. The primary focus is on two kinds of gene-roles: (1) the chromosomal locations of genes, and (2) the biological processes in which genes play a role. Regarding chromosomal locations, the discrepancies revealed are striking and systematic, suggesting a structurally common origin. In regard to the biological processes, difficulties arise because genes frequently play roles in multiple processes, and processes may have many designations (such as synonymous terms). Our algorithms make use of the roles defined in the NCIT Biological Process hierarchy to uncover many probable gene-role errors in the NCIT. These results show that automated comparative auditing is a promising technique that can identify a large number of probable errors and corrections for them in a terminological genomic knowledge repository, thus facilitating its overall maintenance.

1. Introduction

Knowledge of genes and genomes is one of the fastest expanding areas of biomedical research. In 2001, the Human Genome Project (HGP) and the Celera project obtained draft sequences [1,2] of the approximately 3.2 billion nucleotides comprising the human genome. In 2004, the HGP published a complete human DNA sequence. These projects generated a vast body of knowledge from human DNA, including computationally identifying more than 20,000 putative human genes. Obtaining a comprehensive human genome sequence has strongly impacted many areas of biomedical research and medicine [3]. For example, the identification of a disease-related allele of a gene may permit the development of a diagnostic test that reveals a potential health problem before it manifests as symptoms [4]. Knowing a patient's genetic makeup may allow physicians to minimize certain disease risks [5].

Technology plays a critical role in genomic research, which has seen an explosion both of concepts and of data during the last decade. In particular, controlled terminologies [6] are an essential component of this technology. They permit effective access to information in hundreds of genomic databases comprising billions of bytes of information [7]. Terminologies help relate data stored in multiple databases.

The rapid growth of genomic information over the past few years and the nature of the process of its discovery make genomic terminologies particularly susceptible to error. Thus, auditing such terminologies is a major challenge facing the biomedical informatics community.

Because of the important role of genomics in cancer research, the National Cancer Institute Thesaurus (NCIT) [8] provides broad terminological coverage of genomics. Among the genomically related components (hierarchies) of the NCIT are Gene; Gene Product; Biological Process; and Anatomic Structure, System, or Substance. In previous work [9], we applied structural auditing methodologies to the NCIT's Biological Process hierarchy (BPH). In this paper, we develop a comparative auditing methodology to be applied to the NCIT's Gene hierarchy in an effort to find roles (relationships) that are inconsistent and represent possible errors. Whenever we refer to an error identified by our algorithm, we mean such a possible error that is to be brought to the attention of a human auditor. As a side effect, we also audit the BPH for missing concepts and synonyms.

We use the gene databases of the National Center for Biotechnology Information (NCBI), including GenBank [10] and Entrez Gene [11], as the comparative source for the auditing methodology.

* Corresponding author.
E-mail addresses: bcohen@cs.njit.edu, bxcohen@gmail.com (B. Cohen).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.03.010

Both the NCIT and the NCBI provide Web interfaces. We have designed and implemented software to retrieve the corresponding sets of information about human genes from the two sources via the Web. The overall idea is to utilize the knowledge in the NCBI-provided databases to detect missing or erroneous information in the NCIT's Gene and Biological Process hierarchies and to propose corresponding corrections or additions. Naturally, each of the sources is liable to have errors, which could be detected by comparison with the other source. We have chosen in this work to focus on auditing the terminological knowledge source: the NCIT.

One role of a gene recorded in the NCIT on which we utilize our methodology is the gene's location on a chromosome: *Gene_In_Chromosomal_Location*.[1] The comparison of chromosomal locations in the two knowledge sources is relatively straightforward, since each gene has one such target and its range is known in advance. However, the number of discrepancies found with respect to this role is quite striking. These results are reported.

Secondly, we audit the set of target biological processes in which the gene plays a role: *Gene_Plays_Role_in_Process*. Comparison of the knowledge in the two sources in this context is more complicated. A gene typically has a role in multiple biological processes. For each gene, we need to compare the list of target biological processes in the NCBI gene databases with the list of such processes in the NCIT hierarchy. A process in the NCBI list may be identical to or a synonym of a process in the NCIT list—or it may be totally absent from that list. An NCBI process may also be a parent/ancestor or child/descendant (or a synonym thereof) of a process in the NCIT's BPH. One has to detect which one of these cases occurs and then propose the appropriate change. We present an algorithm that considers each of these cases and proposes proper remediation. As a result of the application of our comparative auditing methodology, thousands of potential errors or omissions were discovered in the Gene and Biological Process hierarchies.

Let us note that differences between the knowledge contained in the NCIT and the NCBI regarding a specific gene can arise for many reasons. These include differences in the scientific knowledge sources exploited by the two knowledge-bases and their reliability, different levels of granularity, and outright errors. In this paper, we take the view that our methodology is a tool to aid an auditor in the task of identifying potentially problematic roles of a gene-centered terminology, no matter what the root cause. We refer to all the discrepancies uncovered as "potential errors," since an auditor's responsibility is to raise questions and to suggest potential resolutions. Our methodology complements the suite of tools available to the NCIT's maintenance personnel in the daunting task of assuring the quality of the genomic content of their terminology.

2. Background

2.1. NCI Thesaurus

The National Cancer Institute Thesaurus (NCIT) is a controlled terminology that provides broad coverage of the cancer domain [8,12,13]. It is a public domain terminology that follows a description-logic-based model [14,15].

The NCIT serves as the foundational terminological component in the NCI's efforts to link molecular and clinical cancer-related information [16]. NCIT's array of cancer-related concepts includes cancers themselves, drugs, therapies, genes, biologic processes, proteins, etc. As noted in [16], in response to user needs in the informatics environment, the NCIT's designers and curators have increasingly sought to create a model of how key concepts are defined and relate to each other, thus moving NCIT from a controlled terminology into more of an ontology. The stated goals of its designers make it clear that its intended user group ranges from computer-application developers to researchers to *ad hoc* users [17]. In fact, helping to "speed the introduction of new concepts and new relationships in response to the emerging needs of basic researchers, clinical trials, information services and other users" [17] is one of the primary goals. The NCIT is employed in many NCI applications [17], among them caMOD, the "Cancer Models Database" [18], and caIMAGE, a cancer images database [19]. Among its other current applications, NCIT is being used to aid in the annotations found in the Stanford Tissue Microarray Database (TMAD) [20].

The basic unit of knowledge in the NCIT is the concept. Each concept has a code number and a preferred name (term). Other properties include a definition (English language), a semantic type, and, importantly for our work, a list of synonyms.

The concepts are partitioned into 21 disjoint hierarchies. Each hierarchy consists of a set of concepts linked by IS-A relationships between child and parent, forming a directed acyclic graph. Examples of the hierarchies include Experimental Organism Diagnosis, Biological Process, Gene, and Gene Product.

Roles are directed edges between concepts defining relationships from one to another. These roles can span the different hierarchies. For example, *vegf* is a concept in the Gene hierarchy and *angiogenesis* is a concept in the Biological Process hierarchy. The role *Gene_Plays_Role_in_Process* defines a relationship from *vegf* to *angiogenesis* (see Fig. 1). Roles, also known as associative or semantic relationships, are lateral, in contrast to the hierarchical IS-A relationships. All roles are passed from a parent concept to a child concept via inheritance along the IS-A relationship.

2.2. The NCIT Gene and Biological Process hierarchies

The NCIT editors used HUGO [21] as the authoritative source for their gene concepts because it includes links to a number of other reliable information sources.[2] Entrez Gene [11] and GeneCards [22] were used primarily to validate and to refine chromosomal location data. They also helped in ascertaining processes and diseases. Much of the modeling involved the use of OMIM [23] for disease associations and UniProt [24] for protein functions.

There are 1786 concepts in the Gene hierarchy of the NCIT (2004 version 07.04e). Of these, 1554 are leaves, i.e., concepts having no children. They are the actual gene concepts. There are 232 internal concepts—concepts that have children—that serve to classify the genes into categories. The Gene hierarchy differs from other NCIT hierarchies in that the internal concepts are not themselves gene concepts, just categories. By contrast, an internal concept of the BPH is a biological process. Its children are more refined biological process concepts. Note that nothing is fundamentally wrong with either of these approaches to modeling the internal concepts of a hierarchy. They are just modeling choices made according to differing needs in disjoint hierarchies.

In the BPH, *cancer progression*, for instance, is a biological process that is an internal concept. It has 12 descendant concepts. Some of these, for example, *cancer cell growth*, *metastasis*, and *tumor progression* and its child *tumor expansion*, are also internal concepts. Examples of leaf (childless) biological process concepts include *distant metastasis* and *tumor cell mobility*. Biological process concepts each have only one parent.

[1] Roles and concepts are written in italics.

[2] F. Hartel, personal communication.

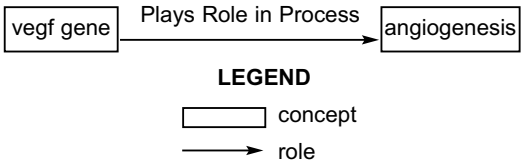

Fig. 1. In the NCIT, a role defines a relation from one concept to another (target) concept.

2.3. The NCBI Gene Databases GenBank and Entrez Gene

The National Center for Biotechnology Information (NCBI) [25] was established in 1988 as a national resource for molecular biology information. The NCBI's GenBank [10,26] is part of the International Nucleotide Sequence Database Collaboration [27]. It is one of three comprehensive repositories of genomic sequence information, along with the EMBL in Europe [28] and the DDBJ in Japan [29]. As of August 2005, GenBank exceeded 100 Gb (billions of nucleotides) of genomic data. GenBank contains, in addition to the sequence data, extensive sequence annotation about each of the complete genomes and each identified or putative gene such as known functions of each gene; links to the original report of the genomic research; gene location within the genome; alternative splicings of each gene; the sequence(s) of the protein product(s), etc. The respective source of each item of information is recorded. GenBank is a primary resource for all kinds of biological, medical, and genomic research.

NCBI's Entrez Gene [11] provides a gene-centered view of the reference sequences contained in the GenBank. It includes published annotations of gene functionality. NCBI's GeneRIF [30] is a mechanism by which scientists and researchers can add to the functional annotations of genes appearing in Entrez Gene. GO codes are often used in this context.

In our methodology, we make use of both the functions and the processes of genes listed in Entrez Gene. These correspond to the biological processes of genes in the NCIT.

This information is accessible via the Web through search and analysis tools, viewable with a variety of visualization tools, and available for public download. We have exploited the Web-posted version to guarantee up-to-date information. Access was accomplished via a Web crawler, discussed below.

2.4. The Gene Ontology (GO)

The Gene Ontology (GO) [31] is used by researchers in reporting their results regarding genes and gene products. Many genomic databases annotate their entries with GO codes, and these are included in the NCBI gene data. GO's evidence codes [32] categorize the sources of genomic knowledge. We extract this information about each biological process target concept to assist a human auditor in evaluating the evidence.

2.5. Ontology comparison and auditing

Auditing is an essential part of terminology and ontology maintenance [9,33], and a great deal of previous work has focused on this issue. In the context of the UMLS, for example, techniques have been developed for the discovery of conceptual ambiguity and redundancy [34] as well as hierarchical circularity [35]. Finding existing concepts that should be merged due to synonymy is an ongoing challenge that has been tackled with various lexical matching methods coupled with phrase substitution [36]. Methodologies addressing a similar synonymy problem (which can include underspecified duplicate concepts) have been developed for auditing description-logic-based terminologies (e.g., SNOMED CT [37], NCIT [8], GALEN [38], and DICE [39]). Algorithms have been devised for finding inconsistencies in such terminologies [40]. In our own previous work, we have employed various abstraction networks (e.g., an object-oriented schema representation [41] and the "Refined Semantic Network" [42]) to glean potentially erroneous concepts within the UMLS.

Other abstraction networks of ours, e.g., the "area taxonomy," have been utilized to audit the NCIT [9]—the main focus of this paper—and SNOMED [43] for various kinds of errors, including redundant concepts, incorrect IS-A arrangements, and erroneous (lateral) relationship configurations. Auditing of the NCIT and SNOMED has been done from the point of view of their adherence to fundamental terminological and ontological principles [44–46]. In particular, SNOMED's IS-A hierarchy has been analyzed in this vein [47]. In [48], such analyses are brought to bear specifically on NCIT's representation of entities pertaining to colon carcinoma.

In the OBO-World [49], of which NCIT is a part, much attention has been paid to GO's alignment with other ontologies and its overall improvement. Since GO has been integrated into the UMLS [7], many of the above mentioned techniques are applicable. In [50], GO was translated into a Protégé [51] ontology and was audited for extraneous IS-A links, for example. Automated means for identifying circular and unintelligible (textual) definitions in GO have been presented in [52]. Various text matching techniques have been employed in an effort to discover relationships between GO concepts and those in other OBO ontologies [53]. It is pointed out in [53] that one needs to be cautious even with perfect matches, which can suffer from polysemy. Caution must be taken with partial matches, e.g., derived using stemming. The fact that standard textual manipulations in the matching process are not well suited for the biomedical domain is also expounded in [54], which, too, deals with the issue of mapping GO and the UMLS. A general fault model for evaluating lexical techniques used in methodologies that perform mapping, alignment, and linking of ontologies (i.e., MALO systems) is presented in [33]. A methodology utilizing both lexical and structural (i.e., involving the hierarchical IS-A and part-whole relationships) methods to align FMA [55] with GALEN appears in [56].

We view the work presented in this paper as straddling the areas of terminology auditing and alignment. We employ structural techniques to correct and enhance the gene-role content of the NCIT. In [57], the use of the "found/fixed graph" metric from the area of software engineering is proposed to assess the quality and completeness of knowledge bases. In this paper, we seek similar assessments, but with an approach that does automated comparison of an existing terminology, the NCIT, with an external knowledge base, namely, the NCBI's Entrez Gene. Our focus is more limited as we concentrate solely on roles of genes. In [58], an external OBO ontology, ChEBI [59], is used to infer additional relationships among GO terms. We are not actually trying to infer such new knowledge in our methodology. We are simply trying to confirm and/or refine the targets (or fillers) of previously defined roles.

An additional unique aspect of our methodology is found in the utilization of one of the NCIT's hierarchies, namely, its BPH, to resolve inconsistencies between another of its hierarchies, the Gene hierarchy, and an outside knowledge source, maintained by the NCBI. Since the target concepts of *Gene_Plays_Role_in_Process* roles of genes all reside in the BPH, it can be used as the basis of comparison between the Gene hierarchy and the NCBI. In particular, it can be used to determine whether discovered discrepancies are really problematic or perhaps just cases of synonym usage. Or its hierarchical configuration can be used to reveal cases of refinement, where a parent concept is the target in one source, whereas its child is the target in the other.

3. Methods

Our methodology seeks to identify possible role errors and omissions in the Gene hierarchy and the BPH of the NCIT by comparing it to the corresponding gene information in another gene knowledge source, namely, Entrez Gene. There are two phases to the approach: information collection and information processing/comparison. Both are carried out algorithmically.

The collection of the desired information from the two knowledge sources was carried out with the use of a pair of Web crawlers. This was done to ensure up-to-date information and avoid any inconsistencies between downloaded local copies of the respective knowledge sources and those versions posted on the Web. The information processing/comparison phase involved the development and implementation of algorithms to identify correspondences and discrepancies between the two knowledge sources, and to recommend remedial action in the case of discrepancies.

3.1. Data collection

The two Web crawlers we developed were used to retrieve data, respectively, from the NCIT and the NCBI via their Web interfaces. Since our target for auditing was the set of human genes common to the two, the Web crawlers were designed to extract the set of human genes and the associated data needed for the auditing process. The relevant gene data was retrieved, parsed, and stored in a relational database. The data retrieved includes a list of the human genes in each source; the list of biological process target concepts for each gene; and the chromosomal location of each gene. The NCIT Web crawler also visited the BPH and identified IS-A relationships among the biological process concepts and their synonyms.

Another set of programs processed the data according to the algorithm described below and displayed the results in tabular form, highlighting the differences between the NCIT and the NCBI and the recommended actions. The first step in the data analysis was to compile a list of genes common to the two knowledge sources for detailed automated comparison.

3.2. Algorithm to audit the NCIT chromosomal location information

For each gene present in both the NCIT and the NCBI, our methodology identifies which of four possible cases applies to the chromosomal location of the gene: (1) information about the chromosomal location may be present in the NCBI and missing from the NCIT; (2) there may be a location in the NCIT and none in the NCBI; (3) the same location may be given in each knowledge base; or (4) different locations may be given in the two knowledge sources. For the purpose of auditing the NCIT, only those genes present in both sources can be processed. If a given gene is only present in the NCIT (or vice versa), then our methodology is not applicable.

3.3. Algorithm to audit the NCIT Biological Process list

Whenever a gene is included in both the NCIT and the NCBI, there are a number of possible cases that must be considered for its biological processes, i.e., the targets of the gene's role *Gene_Plays_Role_in_Process*. In general, a gene may have a role in multiple biological processes. Each process may have synonyms. As noted, information about biological processes is stored in a separate hierarchy of the NCIT: the BPH.

We have developed an algorithm that uncovers and proposes corrections for three types of problems:

(1) An overly general biological process target concept of a gene in the NCIT Gene hierarchy. In this case, it proposes the substitution of a more specific descendant.
(2) A missing biological process of a gene in the NCIT Gene hierarchy. In this case, it proposes the addition of a biological process concept as a target of the gene's *Gene_Plays_Role_in_Process* role.
(3) A biological process target in which a gene plays a role in the NCBI but which is missing from the BPH. In this case, a correction is proposed to both the NCIT's gene concept and its BPH.

Let g be a gene concept that exists in both the NCIT and the NCBI, P_g be the set of biological processes of g in the NCIT, and Q_g be the set of processes and functions of g in the NCBI. Each element q of Q_g (the NCBI biological processes of g) can be used to audit the information in P_g (the NCIT biological processes of g) for completeness and for accuracy. Each q can also be used to audit the BPH for completeness.

This auditing algorithm can be broken down into seven cases. All subsequent references to cases refer to these seven cases. Three of the cases are illustrated in Fig. 2.

Case 1. The simplest case to handle is the one in which the biological process data in the NCBI exactly matches and confirms the NCIT's, using the identical terms. For example, the *cd14* gene exists in both the NCIT and the NCBI. In both repositories, *apoptosis* is listed as a biological process associated with the gene. Since *apoptosis* is among the processes of *cd14* in the NCIT, *apoptosis* must also already be included in the NCIT Biological Process hierarchy. Hence, the audit has found agreement between the two knowledge bases and no corrective action is indicated.

Note that it is possible that an exact term match represents a case of homonymity, where the matched terms have different meanings in the two sources (see, e.g., [33]). This possibility does not concern us here for two reasons. First, the scope of the potential matches is limited specifically to biological processes in both sources. Second, the NCBI constrains our ability to check the real identity of a process term because it has no definitions. So, while the NCIT does contain such information in its BPH, the NCBI's lack of such knowledge makes it impossible to perform any further in-depth comparison.

Case 2. A second case of agreement between the two knowledge bases is one in which a gene has the same associated biological process, but under a synonymous term. For example, the *adra1a* gene has the process *negative regulation of cell proliferation* in the NCBI. It has the process *inhibition of cell proliferation* in the NCIT. Since *negative regulation of cell proliferation* is a synonym of *inhibition of cell proliferation* in the BPH, no corrective action is indicated.

When there is a discrepancy between the biological processes assigned to a gene g in the two knowledge sources, there are three ways that the information about the biological processes of g in the

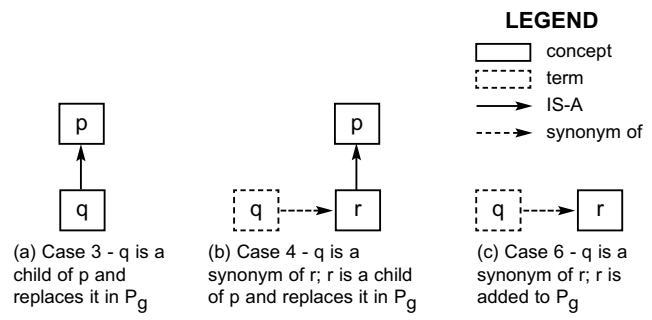

Fig. 2. Three cases of auditing Biological Process role targets of a gene in the NCIT.

NCBI (that is, Q_g) may be used to improve the quality of the information in the NCIT. These are handled, respectively, by Cases 3 and 4 together, Cases 5 and 6 together, and Case 7.

Cases 3 and 4. First, for a particular p in the NCIT, there may be a more specific q in the NCBI (see Fig. 2a). That is, the comparison may provide a q that is a child or other descendant process of p in the BPH. This child process q may be a concept in the BPH (Case 3) or a synonym of a concept in the BPH (Case 4; see Fig. 2b). As an example of Case 3, the *il8* gene has the process *cell proliferation regulation* in the NCIT and *cell cycle arrest* in the NCBI. Since *cell cycle arrest* is a child of *cell proliferation regulation*—that is, a more specific concept—a correction should be made in the NCIT by replacing the existing concept *cell proliferation regulation* with the child concept *cell cycle arrest*. The more specific role implies the more general one (see Fig. 3).

Cases 5 and 6. A second category of discrepancy that the algorithm detects is one in which the NCBI provides a biological process target q for g that is in the BPH but is missing from P_g. In Case 5, the missing target for g exists in the BPH as a concept. In Case 6, the missing target for g exists in the BPH as a synonym of a concept (see Fig. 2c). As an example of Case 5, the gene *tp53bp2* has the role *induction of apoptosis* in the NCBI. *Induction of apoptosis* exists in the BPH, but is not among the roles of *tp53bp2* in the NCIT. The corrective action called for is the addition of *induction of apoptosis* to the roles of *tp53bp2* in the NCIT.

Case 7. The third type of discrepancy is a biological process q assigned to a gene g in the NCBI that is missing both from P_g and from the BPH. For example, the gene *tlr1* has the target *positive regulation of tumor necrosis factor-alpha biosynthesis* in the NCBI. This process does not exist in the BPH. Consequently, a dual corrective action is called for: addition of this new concept to the BPH and its assignment as a target of *tlr1*.[3]

To help substantiate the need for the suggested corrective actions, "evidence codes" gleaned from the NCBI are also provided to the NCIT editors. These evidence codes were included at the behest of the editors.

The algorithm for performing the audit is presented below in pseudocode. The strategy of the algorithm is to iterate through Q_g—the biological process concepts assigned to a gene g in the NCBI—and to consider all the ways that each element q of Q_g may be used to improve or add to the information about g in the NCIT. If a more refined biological process concept is present in Q_g than in P_g, it is substituted. If a biological process concept is present in Q_g but not in P_g, it is added to P_g. If such a biological process concept is missing from the BPH, it is added. Following our remark above, whenever we say "added," we mean "proposed for addition."

Algorithm: *Audit NCIT Gene and Biological Process hierarchies using NCBI Entrez Gene*

Inputs:
Gene hierarchy of the NCIT (GH_NCIT)
Gene database of the NCBI (GD_NCBI)
Biological Process hierarchy of the NCIT (BPH)
Notation:
P_g = the set of biological processes of gene g in GH_NCIT
Q_g = the set of biological processes of gene g in GD_NCBI
p = an element of P_g
q = an element of Q_g

[3] Note that although the new term is a close match to the existing term "Tumor Necrosis Factor," it is narrower and therefore warrants addition as a new term. To see this distinction in GO, see http://amigo.geneontology.org/cgibin/amigo/go.cgi?view=details&search_constraint=terms&depth=0&query=GO:0042535.

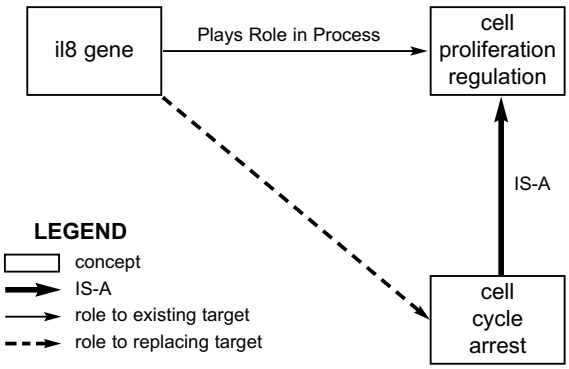

Fig. 3. The more refined concept replaces the more general concept as the target of the role.

```
r = a concept in the BPH
g = a gene that is in both GH_NCIT and GD_NCBI
for each g
   for each q
      //Case 1: q is in the biological process list of g
         if q ==p for some p in P_g
            do: nothing
      //Case 2: q is a synonym of a biological process of g
         else if q is a synonym in BPH of some p in P_g
            do: nothing
      //Case 3: q is a descendant of a biological process of g
      //(Fig. 2a)
         else if q is a descendant of p in BPH for some p in P_g
            do: replace p with q
      //Case 4: q is a synonym of a descendant of a biological
      //process of g (Fig. 2b)
         else if q is a synonym of r
            and r is a descendant in BPH of p for some p in P_g
         do: replace p with r
      //Case 5: q exists in BPH but is missing from P_g
         else if q is in BPH
            do: add q to P_g
      //Case 6: q is a synonym of a concept in BPH that is
      //missing from P_g
         else if q is a synonym of concept r in BPH
            do: add r to P_g
      //Case 7: q is missing from BPH and thus from P_g
         else
            do: add q to the BPH
            do: add q to P_g
```

This pseudocode presentation of the algorithm is designed for clarity and simplicity, and abstracts from considerations of efficiency of the implementation, particularly in the identification of hierarchical relationships. A number of optimizations of the implementation are available. In practice, the algorithm can be executed on a desktop computer in a matter of seconds, even on large data sets.

4. Results

4.1. Chromosomal Location role

Our comparison revealed an unexpectedly large number of differences between the chromosomal locations found in the NCIT and the NCBI. Of the 1960 genes examined, 576 (29.4%) displayed differences between the NCBI and the NCIT. In 262 instances, a chromosomal location was present in the NCBI but missing from the NCIT. For example, *adamts1* gene has location 21*q*21.2 in the NCBI, but none in the NCIT. In 171 instances, there was a location

in the NCIT but no location at all in the NCBI. As an example, *kras2* gene has location 2*p*12.1 in the NCIT, but none in the NCBI. Finally, in 143 instances, a location was given in each knowledge source, but they differed. The location, for example, of the *ereg* gene is given as 4*q*13.3 in the NCBI and as 4*q*21.1 in the NCIT.

4.2. Biological Process role

There are 1462 genes in both the NCIT hierarchy and the NCBI for which the NCBI identifies at least one biological process as a target of the respective "plays role in process" role. (Of the total 1960 genes examined, 498 were without any biological process target in the NCBI.) For these genes, the NCBI has 1802 distinct associated biological processes, while the NCIT has 287. The total numbers of associated biological processes for the genes are 10,902 and 4597 in the NCBI and the NCIT, respectively. A biological process may be associated with multiple genes in either knowledge source. For example, six genes in the NCBI, namely, *clic4*, *gria3*, *itpr1*, *slc5a5*, *stim1*, and *tf*, play a role in the biological process *ion transport*. Similarly, a gene may play a role in multiple biological processes. The gene *cdc37* plays a role in four processes, *protein folding*, *protein targeting*, *regulation of cyclin-dependent protein kinase activity*, and *unfolded protein binding*, in the NCBI and four processes, *signal transduction*, *subcellular protein targeting*, *protein folding*, and *cell cycle regulation*, in the NCIT.

The comparison of biological process targets between the NCIT and the NCBI identified numerous candidates for modification. Since the results are too voluminous to be presented here in full, for each case considered by the algorithm, we present a sample of the results. In the remainder of this section, whenever we refer to a "target," we mean "biological process target."

Case 1. The biological process in the NCIT is the same as that in the NCBI. No action is necessary. We found 602 such targets involving 67 distinct biological processes. A sample is contained in Table 1. For example, the gene *vegf* has the target *angiogenesis* in both sources.

Case 2. The target in the NCBI is a synonym of one in the NCIT. No action is necessary. We found 40 such targets involving six distinct biological processes. As an example, for gene *f*9, the NCBI target *blood coagulation* is a synonym in the BPH of the NCIT target *coagulation*. A sample is contained in Table 2.

Case 3. The target in the NCBI is a descendant in the BPH (that is, a refinement) of a target in the NCIT. The algorithm recommends the replacement of the more general concept with the more specific one. We found 22 such targets involving 13 distinct biological processes in the NCBI and nine distinct biological processes in the NCIT. For example, the gene *cdkn1b* has the target *cell cycle arrest* in the NCBI and *cell cycle regulation* in the NCIT. *Cell cycle arrest* is a child of *cell cycle regulation* in the BPH. These cases are contained in Table 3.

Case 4. A target in the NCBI is a synonym of a descendant in the BPH of a target of the same gene in the NCIT. The algorithm recommends replacing the more general concept with the more specific one. We found only one instance of this: the gene *pthlh* has the target *negative regulation of cell proliferation* in the NCBI and the target *cell proliferation regulation* in the NCIT. It is recommended that the target be changed to the concept *inhibition of cell proliferation*, the synonym of which in BPH is *negative regulation of cell proliferation*.

Case 5. A biological process that exists in the BPH is a target in the NCBI but not in the NCIT. The algorithm recommends that that target be added to the NCIT. For example, the *trex1* gene has the target *mismatch repair* in the NCBI. *Mismatch repair* is a biological process in the BPH, but is not a target for *trex1* in the NCIT. Therefore, *mismatch repair* is recommended for addition to *trex1* in the NCIT. We found 1694 such instances, involving 87 distinct biological process concepts. A sample of these appears in Table 4.

Table 1
Sample of identical biological process targets in the NCBI and the NCIT

Gene	Target
vegf	Angiogenesis
app	Apoptosis
bax	Apoptosis
il10	B Cell proliferation
insr	Carbohydrate metabolism
app	Cell adhesion
cdc6	Cell cycle
cdkn2b	Cell cycle arrest
brca1	Cell cycle checkpoint
bmp3	Cell differentiation
ppp1ca	Cell division
fn1	Cell migration
hmmr	Cell motility
fgf1	Cell proliferation
cd44	Cell–matrix adhesion

Table 2
Sample of synonymous targets in the NCBI and the NCIT

Gene	NCBI target	NCIT target
f9	Blood coagulation	Coagulation
ins	Cell–cell signaling	Intercellular communication
mcm5	DNA replication initiation	Replication initiation
phb	Negative regulation of cell proliferation	Inhibition of cell proliferation
ercc4	Nucleotide-excision repair	Nucleotide excision repair
bmp10	Regulation of cell proliferation	Cell proliferation regulation

Table 3
A target in the NCBI is a descendant of a target in the NCIT

Gene	Descendent role in NCBI	Role in the NCIT
cdkn1b	Cell cycle arrest	Cell cycle regulation
cdkn1c	Cell cycle arrest	Cell cycle regulation
inha	Cell cycle arrest	Cell proliferation regulation
ccne2	Cell cycle checkpoint	Cell cycle regulation
cenpf	Cell division	Cell division process
cd22	Cell–cell adhesion	Cell adhesion
icam1	Cell–cell adhesion	Cell adhesion
itgb4	Cell–matrix adhesion	Cell adhesion
hk2	Glycolysis	Carbohydrate metabolism
gnas	GTP binding	Ligand binding
alox15	Inflammatory response	Inflammation process
icam2	Integrin binding	Ligand binding
tgfbi	Integrin binding	Ligand binding
pms2l1	Mismatch repair	DNA repair
cenpf	Mitosis	Cell division process
pik4cb	Receptor mediated endocytosis	Endocytosis

Table 4
A biological process in the BPH is a target in the NCBI but not in the NCIT

Gene	Role in the NCBI
aim2	Immune response
col4a3	Induction of apoptosis
inha	Induction of apoptosis
rela	Inflammatory response
erbb2ip	Integrin binding
ptgs1	Keratinocyte differentiation
dlg4	Learning
fyn	Learning
cetp	Lipid binding
cyp21a2	Lipid binding
ldlr	Lipid transport
shh	Lung development
rb1	m Phase
exo1	Meiosis
msh4	Meiosis
mlh3	Meiotic recombination

Case 6. A target assigned to a gene in the NCBI is a synonym of a process in the BPH and is not a target for the same gene in the NCIT. We found no instances of this case.

Case 7. A biological process target in the NCBI is not in the BPH. The algorithm recommends that that target be added to the gene in the NCIT and the biological process be added to the BPH. For example, the gene *anxa5* plays a role in the biological process *anti-apoptosis* in the NCBI. *Anti-apoptosis* is not in the BPH, and therefore is also not a target of *anxa5* in the NCIT. A*nti-apoptosis* is recommended for addition to the BPH and to the targets of *anxa5* in the NCIT. We found 10,584 such instances, involving 1611 distinct biological process concepts. A sample of these is shown in Table 5. Let us note that the number of concepts involved is larger than the current number of concepts in the BPH.

4.3. Evidence code usage

All roles in the NCBI Entrez Gene are annotated with one or more of the 14 GO evidence codes [11]. As noted, the evidence codes are provided to assist a human auditor in deciding on the disposition of an algorithmically generated proposed action. The codes and their frequency of occurrence in the genes analyzed in the NCBI are listed in Table 6. For example, the code IC (Inferred by Curator) occurred 13 times. Most targets are annotated with a single code. In 510 instances, targets were assigned to genes based on multiple evidence codes.

Table 5
Targets of a gene in the NCBI that are not in the BPH

Gene	Role in the NCBI
ins	Alpha-beta T cell activation
vcl	Alpha-catenin binding
tat	Amino acid and derivative metabolism
bphl	Amino acid and derivative metabolism
cad	Amino acid binding
ar	Androgen binding
shbg	Androgen binding
cyp11a1	Androgen biosynthesis
ar	Androgen receptor activity
ar	Androgen receptor activity
daxx	Androgen receptor binding
brca1	Androgen receptor binding
bcl2	Anti-apoptosis
cdc2	Anti-apoptosis
ppp2r1b	Antigen binding
il7r	Antigen binding
mt3	Antioxidant activity
slc26a3	Antiporter activity
vcl	Apical junction assembly
vdac1	Apoptogenic cytochrome *c* release channel activity
top2a	Apoptotic chromosome condensation
bak1	Apoptotic mitochondrial changes
bax	Apoptotic mitochondrial changes
bad	Apoptotic program
bik	Apoptotic program

Table 6
GO evidence codes annotating targets in the NCBI, and their frequency

Evidence code	Evidence code meaning	Instances
IC	Inferred by curator	13
IDA	Inferred from direct assay	718
IEA	Inferred from electronic annotation	4928
IEP	Inferred from expression pattern	42
IGC	Inferred from genomic context	0
IGI	Inferred from genetic interaction	5
IMP	Inferred from mutant phenotype	122
IPI	Inferred from physical interaction	520
ISS	Inferred from sequence or structural similarity	494
NAS	Non-traceable author statement	851
ND	No biological data available	73
RCA	Inferred from reviewed computational analysis	0
TAS	Traceable author statement	3415
NR	Not recorded	0

5. Discussion

Our algorithm is novel in its use of terminology relationships such as parent, child, and synonym in one hierarchy—the Biological Process hierarchy of the NCIT–in conjunction with screening knowledge in a different biomedical resource—the Entrez Gene database of the NCBI—to audit another hierarchy—the Gene hierarchy of the NCIT.

A large number of possible errors were discovered by our automated means. With regard to the chromosomal location role, the data discrepancies that we uncovered were quite striking, suggesting some systematic origin. Our methodology is intended to identify such significant discrepancies between the genomic knowledge sources, but it does not seek to provide for their resolution. Database maintainers whom we queried could not provide an immediate explanation. Among the possible explanations are the confusion of multiple copies of a gene and the disparities between multiple sequencings. Our finding regarding discrepancies in chromosomal location is verified by the independent update in the current (07.12e) version of the NCIT, in which 117 of the 136 genes that previously had incorrect chromosomal locations now have corrected locations that are the same as our suggestions.

Instances of six of the seven possible types of biological process errors were discovered. The number of instances of such errors varied from one to thousands. The largest number involved biological processes that are identified in the NCBI but are not included in the NCIT's BPH. The documentation of such a large number of possible errors or omissions through comparative auditing should assist human experts in significantly improving the quality, and especially the consistency of coverage, of gene-roles.

A sample of these results was submitted for review to the team maintaining the NCIT. Expert review is important in assessing the usefulness of the results in improving the efficiency of the work of the domain experts.

We believe that an automated process may be particularly useful in assisting a human auditor in elucidating distant relationships. For example, the biological process *cell cycle arrest* is a role in the NCBI of the gene *gadd45a*. *Cell cycle arrest* is a child in the BPH of the biological process *cell cycle inhibition*, which in turn is a child of *cell cycle regulation*, which is assigned to *gadd45a* in the NCIT. That is, the role assigned to *gadd45a* in the NCBI is the grandchild of its assigned role in the NCIT. The role *cell cycle regulation* is also assigned to genes *cdkn1b*, *cdkn1c*, and *cdkn2c* in the NCIT. The same biological process *cell cycle arrest* is a target in the NCBI of the gene *imha*. The gene *imha* has the target *cell proliferation regulation* in the NCIT, which is the parent of *cell cycle regulation*. That is, the role assigned to *imha* in the NCBI is the great grandchild of the role assigned to it in the NCIT. In each of these cases, the algorithm recommends the replacement of the ancestral process with the descendant one.

5.1. Possible extensions of automated comparison

The automated methods we employed could be extended in several ways. In our algorithm, only the genes that appear in both the NCIT and the NCBI are examined. Of course, the two knowledge sources can be compared to discover gene concepts that appear in one but not the other.

One-way auditing—using the NCBI to audit the NCIT—could readily be extended to two-way auditing. The information P_g about gene g in the NCIT could, in principle, be used to audit the information Q_g about gene g in the NCBI. That would be fruitful when the

information P_g is more complete or more specific than that in Q_g. Parallel auditing in the opposite direction would require only a straightforward, symmetric extension of the auditing algorithm. For example, for some gene g that exists in the NCBI and the NCIT, there could be a biological process p in the NCIT that is a descendant of some biological process q in the NCBI and which should therefore be substituted for it. There could also be a p that is missing from Q_g and should be added to it. To illustrate this, Table 7 gives some examples derived by manual inspection. NCIT biological process targets that are assigned to genes in the NCIT but not in the NCBI are in bold; NCIT biological process targets that are apparently more refined and their corresponding targets in the NCBI are in bold italic.

Our algorithm audits the Gene hierarchy and the Biological Process hierarchy of the NCIT using the Entrez Gene database of the NCBI. One could extend this method by utilizing yet other terminologies, as suggested by an anonymous reviewer. A major obvious resource is the GO terminology, which is widely used for annotation. One could also extend the straightforward string matching that we used by incorporating other lexical matching techniques, for example, normalization.

5.2. Limitations

The comparative auditing methodology applied here can detect problems and suggest resolutions, but cannot definitively resolve discrepancies. The subtle evaluation of meaning by a human reviewer is still required to resolve the proposed changes. Auditing a terminology is labor-intensive and the time of domain-expert auditors is limited. The purpose of the automated methodology is to achieve the most productive and accurate use of the experts by identifying possible errors and offering potential resolutions.

Some types of problems continue to escape machine detection, and in some cases the algorithm makes a faulty recommendation. We subjected a subset of the algorithm's recommendations to a manual review by one of the authors (HM) with medical training. Specifically, the output of the algorithm for a sample of 100 genes was reviewed. For these 100 genes, there were a total of 812 associated occurrences of processes. The algorithm confirmed 55 processes as being valid in the NCIT, and reported that no action was required. These were the cases where the NCBI process associated with the specific gene was either identical to, a synonym of, or a parent of the NCIT process (in the BPH). For the rest of the 762 processes, the manual review of the corrective actions suggested by the algorithm revealed that 89 of them (about 12%) were, in fact, incorrect and should not be acted upon.

In some cases, the algorithm failed to detect synonymy of biological functions. This happened when the strings representing the functions were not identical and the different strings were not identified as synonyms in the BPH. For example, the *mas1* gene has the biological function *G-protein coupled receptor protein signaling pathway* in the NCBI and the biological process *G-protein coupled receptor signaling* in the NCIT. Since these two strings are not identical and are not listed as synonyms in the BPH, the algorithm believed that *G-protein coupled receptor protein signaling pathway* was a new biological process for the NCIT and recommended that it be added as a target of the *mas1* gene and that it be added to the BPH. The human reviewer detected that these two processes were synonymous. No action, in fact, is required. The human reviewer's observation, while overriding the algorithm's faulty recommendation, leads to another type of error to correct: *G-protein coupled receptor protein signaling pathway* should be identified in the BPH as a synonym of *G-protein coupled receptor signaling*. This illustrates another feature of auditing: the correction of a single error often propagates to other instances of the same error. In this case, once the human editor finds this error made by the algorithm, the same correction could be applied to 25 other genes. Those occurrences can be detected by a simple string matching search over the algorithm's results.

In some cases, the algorithm failed to detect parent–child relationships between biological functions in the two knowledge sources. This happened, for example, when an NCBI target for a gene was not properly detected as a parent of the NCIT target. For example, the *lats2* gene has the biological process *G-protein amino acid phosphorylation* in the NCBI and the biological process *serine–threonine phosphorylation* in the NCIT. Because *protein amino acid phosphorylation* is not in the BPH, the algorithm did not detect that *protein amino acid phosphorylation* is a parent concept of *serine–threonine phosphorylation*. It therefore recommended the addition of *protein amino acid phosphorylation* to both *lats2* gene and the BPH. The human reviewer detected that these two processes are related, and that the more refined concept is the one listed for *lats2* gene in the NCIT. Again, the human reviewer's observation overrides the algorithm's recommendation and leads to another type of error to correct: *protein amino acid phosphorylation* should be added to the BPH as a synonym of *protein phosphorylation*, the parent of *serine–threonine phosphorylation*.

Another class of error detected by the human reviewer involves a target in the NCBI that should be added to the BPH as a child of an existing process (Table 8). For example, among the NCBI targets are

Table 7
Possible errors in the NCBI

Gene	NCBI Biological Process target	NCIT Biological Process target
tf	Ion transport, iron ion homeostasis, ron ion transport	**Ligand binding**, transport process, metal ion binding, **immune function**
adamts1	Integrin-mediated signaling pathway, negative regulation of cell proliferation, proteolysis	**Angiogenic inhibition**, inhibition of cell proliferation, proteolysis
rbm6	***RNA processing***	**Tumor suppression, ligand binding**, ***RNA binding***
plau	***Blood coagulation***, chemotaxis, fibrinolysis, proteolysis, proteolysis, signal transduction	Proteolysis, ***anticoagulation***
igf2	***Cell proliferation***, development, imprinting, insulin receptor signaling pathway, physiological process, regulation of progression through cell cycle, skeletal development	***Stimulation of cell proliferation***, intercellular communication
slc5a5	Ion transport, sodium ion transport	**Drug efflux**, transport process, **multidrug resistance, ligand binding**
serpinb5	Cell motility	**Cell–matrix adhesion, ligand binding, metastasis suppression, tumor suppression**
selp	***Cell adhesion***	***Cell–matrix adhesion, immune function, ligand binding***

Targets missing in the NCBI are in bold; more refined targets in the NCIT and the corresponding ones in the NCBI are in bold italic.

Table 8
NCBI biological processes that are detected by human review as refinements of the BPH concept *transport process*

Gene	Refined role in the NCBI
slc26a3	Anion transport
fyn	Calcium ion transport
ccl8	Calcium ion transport
tf	Iron ion transport
tfr2	Iron ion transport
slc5a5	Sodium ion transport
slc26a3	Sulfate transport

several refinements of the BPH concept *transport process*, including *anion transport*, *calcium ion transport*, *iron ion transport*, *sodium ion transport*, and *sulfate transport*. In this set of errors, human comprehension was critical in discerning that the concept *transport* could be refined by the addition of a number of qualifying prefixes.

The human reviewer noted that the NCBI targets sometimes contain redundancies, with one target for a given gene being a more refined version of another target. In this case, the correct action is to add only the more refined concept, if it is not already among the NCIT targets for the given gene. The more general concept should not be added. For example, in the NCBI, the gene *bmp10* is identified as playing a role in *embryonic development* and in *embryonic heart tube development*. The latter is a child of the former. Table 9 gives a sample of this type of problem. We note that in many cases, as in all cases of Table 9, the words of the general process appear in the name of the more refined process. Potentially such cases could be detected automatically by string matching.

The human reviewer also noted that the NCBI provides two strings (*molecular function unknown* and *biological process unknown*) as null indicators (73 instances).

The algorithm we developed for auditing, based on discrepancies in the biological process information of genes, makes certain simplifying assumptions. We assume that no two biological process roles of a gene in the NCIT have an ancestor–descendant relationship to each other and that there are no errors in the BPH except for possible omissions. Had we extended the algorithm to cover cases excluded by these assumptions, additional errors might have been discovered. However, the simplifying assumptions do not invalidate any of the reported results.

Finally, we note that the NCIT editor reviewing the results of the automatic auditing algorithm may consider the GO evidence code (see Table 6) for each recommendation provided by the algorithm. While the NCBI stores knowledge from various publications (with the proper evidence code), the NCIT requires some level of reliability of the evidence to justify its inclusion.[4] This explains some of the large differences in the listings in the two knowledge sources. As a matter of fact, the NCIT editors can automate the elimination of recommendations with low-level evidence codes and thus avoid their manual review.

Table 9
Redundant targets due to parent–child relations in the NCBI

Gene	Redundant NCBI Biological Process role	NCBI Biological Process role
bmp10	Embryonic development	Embryonic heart tube development
il7	Positive regulation of cell proliferation	Positive regulation of B cell proliferation
chrm4	Receptor activity	Rhodopsin-like receptor activity
tgfbr1	Transforming growth factor beta receptor activity	Transforming growth factor beta receptor activity type i
gtf3c4	Transferase activity	Acyltransferase activity
ptk2	Binding	ATP binding
adrb1	Receptor activity	Beta1-adrenergic receptor activity
ltb4r	Muscle contraction	Cardiac muscle contraction
ccbp2	Receptor activity	c–c Chemokine receptor activity
lox	Metal ion binding	Copper ion binding
blr1	Receptor activity	c–x–c Chemokine receptor activity
emp1	Development	Epidermis development

6. Conclusions

We have presented an automated auditing methodology focused on role errors in the Gene hierarchy of the NCIT. Our methodology utilizes the NCBI's Entrez Gene for comparative review. Pertinent information is extracted by two Web crawlers. The results are then computationally classified, using algorithms developed for each of two roles: *Gene_In_Chromosomal_Location* and *Gene_Plays_Role_in_Process*.

Our results show that the very rapid accumulation of genomic data and its storage in ontologies and other knowledge-bases has been accompanied, as one could expect, by a significant error rate. The greatest source of error encountered is the lag in incorporating and updating the data, but other ontological problems are also significant. The rate of accumulation of genomic knowledge is, furthermore, continuing to increase. Since the availability of human domain experts is limited, this implies a daunting volume of errors that threatens to overwhelm the human resources available to curate and audit genomic knowledge databases. Uncorrected errors in terminologies may propagate to genomic research, impacting on the conclusions researchers draw. Automated methods, including the comparative methodology reported here, show the promise of wholesale attacks on entire classes of errors that scale with the volume of genomic knowledge.

Our comparative auditing techniques are advantageous in that the methods are formalized, reproducible, and scalable. The computational requirements for bandwidth, storage, and computing power are sufficiently modest that they do not constrain the application of these methods.

However, the errors detected are limited to certain categories that an algorithm can identify. The algorithm may misidentify errors, but in the process it can give hints of other errors to a human auditor. Utilizing automatic auditing can make the work of a human editor more efficient, better utilizing the limited auditing resources available.

Acknowledgments

This work was partially supported by the United States National Library of Medicine under Grant R-01-LM008445-01A2. We thank Frank Hartel and Larry Wright from NCI for their assistance and for points of clarification they provided during this research.

References

[1] The International Human Genome Mapping Consortium. A physical map of the human genome. Nature 2001;409:934–41.
[2] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291:1304–51.
[3] Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L, et al. New goals for the U.S. Human Genome Project: 1998–2003. Science 1998;282:682–9.
[4] Lin JH. Divining and altering the future: implications from the Human Genome Project. Science 1998;282:1532.
[5] Karanjawala ZE, Collins FS. Genetics in the context of medical practice. JAMA 1998;280(17):1533–4.
[6] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. JAMIA 1994;1(1):35–50.
[7] Lomax J, McCray AT. Mapping the Gene Ontology into the Unified Medical Language System. Comparative and Functional Genomics 2004;5(5):354–61.
[8] Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. Journal of Biomedical Informatics 2005;38(2):114–29.
[9] Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. JAMIA 2006;13(6):676–90.
[10] GenBank Overview. <http://www.ncbi.nlm.nih.gov/Genbank/index.html/>; 2008 [accessed January 8].
[11] Entrez Gene. <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene/>; 2008 [accessed January 8].
[12] National Cancer Institute Terminology Resources. <http://www.nci.nih.gov/cancerinfo/terminologyresources/>; 2008 [accessed January 8].
[13] NCICB: caCORE. <http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/>; 2008 [accessed January 8].
[14] Nardi D, Brachman RJ. An introduction to description logics. In: Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF, editors. The description logic handbook: theory, implementation, and applications. Cambridge, UK: Cambridge University Press; 2003. p. 1–40.

[4] F. Hartel, personal communication.

[15] Brachman RJ, Schmolze J. An overview of the KL-ONE knowledge representation system. Cognitive Science 1985;9(2):171–216.
[16] Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics 2007;40(1):30–43.
[17] de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI Thesaurus: using science-based terminology to integrate cancer research results. In: Fieschi M, Coiera E, Li YC, editors. Proceedings of Medinfo2004. San Francisco, CA; 2004. p. 33–37.
[18] caMOD—Cancer Models Database. <http://cancermodels.nci.nih.gov/>; 2007 [accessed December 26].
[19] caIMAGE—Cancer Images Database. http://cancerimages.nci.nih.gov/caIMAGE; 2007 [accessed December 26].
[20] Shah NH, Rubin DL, Espinosa I, Montgomery K, Musen MA. Annotation and query of tissue microarray data using the NCI Thesaurus. BMC Bioinformatics 2007;8:296.
[21] HUGO Gene Nomenclature Committee. <http://www.genenames.org/>; 2008 [accessed January 17].
[22] GeneCards. <http://www.genecards.org/>; 2008 [accessed January 17].
[23] OMIM: Online Mendelian Inheritance in Man. <http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim/>; 2008 [accessed January 17].
[24] UniProt: The Universal Protein Resource. <http://www.pir.uniprot.org/>; 2008 [accessed January 17].
[25] National Center for Biotechnology Information, Available from: <http://www.ncbi.nlm.nih.gov/>.
[26] Wheeler D et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 2007;35:D5–D12. doi:10.1093/nar/gkl1031.
[27] INSDC: International Nucleotide Sequence Database Collaboration. <http://www.insdc.org/>; 2008 [accessed January 8].
[28] European Molecular Biology Laboratory. Available from: http://www.ebi.ac.uk/embl/>.
[29] DNA Databank of Japan. Available from: <http://www.ddbj.nig.ac.jp/>.
[30] GeneRIF—Gene Reference Into Function. <http://www.ncbi.nlm.nih.gov/projects/GeneRIF/>; 2008 [accessed January 8].
[31] Gene Ontology Consortium. Creating the Gene Ontology resource: design and implementation. Genome Research 2001;11:1425–433.
[32] Guide to GO Evidence Codes. <http://www.geneontology.org/GO.evidence.shtml/>; 2008 [accessed January 8].
[33] Johnson HL, Bretonnel Cohen K, Hunter L. A fault model for ontology mapping, alignment, and linking systems. In: Proceedings of the pacific symposium on biocomputing 12. Maui, HI; 2007. p. 233–244.
[34] Cimino JJ. Battling Scylla and Charybdis: the search for redundancy and ambiguity in the 2001 UMLS Metathesaurus. In: Bakken S, editor. Proceedings of the 2001 AMIA annual symposium. Washington, DC; 2001. p. 120–24.
[35] Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. In: Bakken S, editor. Proceedings of the 2001 AMIA annual symposium. Washington, DC; 2001. p. 57–61.
[36] Hole WT, Srinivasan S. Discovering missed synonymy in a large concept-oriented Metathesaurus. In: Overhage JM, editor. Proceedings of the 2000 AMIA annual symposium. Los Angeles, CA; 2000. p. 354–58.
[37] IHTSDO: SNOMED CT. <http://www.ihtsdo.org/our-standards/snomed-ct/>; 2007 [accessed December 31].
[38] Rector A, Bechhofer S, Goble C, Horrocks I, Nowlan W, Solomon W. The GRAIL concept modeling language for medical terminology. Artificial Intelligence in Medicine 1997;9:139–71.
[39] De Keizer NF, Abu-Hanna A, Cornet R, Zwersloot-Schonk JH, Stoutenbeek CP. Analysis and design of an ontology for intensive care diagnoses. Methods of Information in Medicine 1999;38(2):102–12.
[40] Schlobach S, Huang Z, Cornet R, Van Harmelen F. Debugging incoherent terminologies. Journal of Automated Reasoning 2007;39:317–49.
[41] Gu H, Perl Y, Geller J, Halper M, Liu L, Cimino JJ. Representing the UMLS as an OODB: modeling issues and advantages. JAMIA 2000;7(1):66–80. Selected for reprint in: Haux R, Kulikowski C, editors. Yearbook of medical informatics: digital libraries and medicine (International Medical Informatics Association). Stuttgart, Germany: Schattauer; 2001. p. 271–285.
[42] Geller J, Gu H, Perl Y, Halper M. Semantic refinement and error correction in large terminological knowledge bases. Data & Knowledge Engineering 2003;45(1):1–32.
[43] Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural methodologies for auditing SNOMED. Journal of Biomedical Informatics 2007;40(5):561–81.
[44] Ceusters W, Smith B, Kumar A, Dhaen C. Mistakes in medical ontologies: where do they come from and how can they be detected? In: Pisanelli DM, editor. Ontologies in medicine: proceedings of the workshop on medical ontologies. Rome; 2003. p. 145–64.
[45] Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. In: Fieschi M, Coiera E, Li YC, editors. Proceedings of Medinfo 2004. San Francisco, CA; 2004. p. 482–86.
[46] Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods of Information in Medicine 2005;44:498–507.
[47] Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in DL-based terminologies: a case study in SNOMED CT. In: Hahn U, Schulz S, Cornet R, editors. Proceedings of the first international workshop on formal biomedical knowledge representation (KR-MED 2004). Whistler, Canada; 2004. p. 12–20.
[48] Kumar A, Smith B. Oncology ontology in the NCI Thesaurus. In: AIME 2005 (Artificial Intelligence in Medicine Europe). Lecture notes in computer science, 3581. Heidelberg, Germany: Springer-Verlag; 2005. 213–20.
[49] The Open Biomedical Ontologies. <http://www.obofoundry.org/>; 2008 [accessed January 2].
[50] Yeh I, Karp PD, Noy NF, Altman RB. Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). Bioinformatics 2003;19(2):241–8.
[51] The Protégé Ontology Editor and Knowledge Acquisition System. <http://protege.stanford.edu/>; 2008 [accessed January 2].
[52] Köhler J, Munn K, Rüegg A, Skusa A, Smith B. Quality control for terms and definitions in ontologies and taxonomies. BMC Bioinformatics 2006;7:212.
[53] Johnson HL, Bretonnel Cohen K, Baumgartner WA, Lu Z, Bada M, Kester T, et al. Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. In: Altman RB, Dunker AK, Hunter L, Murray TA, Klein TE, editors. Proceedings of the pacific symposium on biocomputing 11. Maui, HI; 2006. p. 28–39.
[54] Sarkar IN, Cantor MN, Gelman R, Hartel F, Lussier YA. Linking biomedical information and knowledge resources: GO and UMLS. In: Altman RB, Dunker AK, Hunter L, Jung T, Klein TE, editors. Proceedings of the pacific symposium on biocomputing 8. Lihue, HI; 2003. p. 439–450.
[55] Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. Journal of Biomedical Informatics 2003;36(6):478–500.
[56] Zhang S, Bodenreider O. Aligning representations of anatomy using lexical and structural methods. In: Musen MA, editor. Proceedings of the 2003 AMIA annual symposium. Washington, DC; 2003. p. 753–57.
[57] Baumgartner WA, Bretonnel Cohen K, Fox LM, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics 2007;23(13):i41–8.
[58] Burgun A, Bodenreider O. An ontology of chemical entities helps identify dependence relations among Gene Ontology terms. In: Hahn U, Valencia A, editors. Proceedings of the first international symposium on semantic mining in biomedicine (SMBM-2005). Hinxton, UK; 2005. http://ceur-ws.org/Vol-148; 2007 [accessed December 28].
[59] Chemical entities of biological interest. Available from: [49]; 2008 [accessed January 2].